Supercomputing Detection of Swanson's Relationship Between Raynaud's Disease and Dietary Fish Oil

GSLIS technical report number: ISRN UIUCLIS--2002/2+UPK

Dated: 23 January 2002

Author: Larry S. Jackson, http://people.lis.illinois.edu/~lsjackso/

This report is available on the web.

Abstract

Utilizing scientific literature as a potential source for new knowledge is an extremely attractive idea. Swanson has repeatedly demonstrated that complementary pairings of ideas from existing knowledge domains can be applied in the solution of new problems. However, the nature of the information search used to identify such pairings is fundamentally different from the searches of databases or digital libraries. Pairings exhibiting direct interrelationships in their literatures are presumably already known to be related, and to be retrievable via existing techniques. Relationships must be found that would currently escape detection, such as indirect relationships via some third knowledge domain.

Supercomputers can perform portions of the indirect search process more quickly. They also make possible, indeed, require, a structurally different approach. The literatures of two research topics are tentatively assumed to be related. The search for avenues of relatedness may then proceed from "both ends toward the middle". Reported herein, an attempt along these lines has proven successful in replicating detection of the indirect relationship between Raynaud's Disease and dietary fish oil, as originally discovered by Swanson using largely manual techniques.

The supercomputing analytical approach reduces reliance upon human expertise in the early, screening stages. As such, it can be "scaled up" more readily, enabling at least a cursory statistical examination of many more pairings of research literatures. Not arguing for the substitution of relatively weak statistical clues in preference to researcher understanding in the search process, it is possible to employ the supercomputer as a form of filter or sorter. Research literature pairs exhibiting stronger than normal statistical ties could be identified, based on the joint occurrence of certain highly ranked tokens. Such results could be used to suggest research directions, or to help in selecting among choices of research

options.

Overview of Undiscovered Public Knowledge

Utilizing scientific literature as a potential source for new knowledge is an extremely attractive idea. Beginning with [1] in 1986, Don R. Swanson demonstrated complementary pairings of items of existing knowledge, extracted from the public literature, could be successfully applied in the solution of new problems. Swanson showed the existence of an indirect statistical link between the literature of Raynaud's disease (dating back to the publication of [2] in 1862) and the literature concerning dietary fish oil that implied a possible amelioration of the symptoms of the disease. This ameliorative effect was subsequently confirmed medically [3]. Swanson similarly detected additional unknown synergistic pairings of literatures addressing other medical problems [4]. This paper retraces the statistical linkages between the literatures concerning Raynaud's Disease and dietary fish oil, but using a supercomputing-based approach. Detection of the statistical linkage is shown to be possible using this approach as well.

It is certainly plausible that a new problem could be solvable through combination of two or more extant research results. But, if the researchers involved are unaware of one another's work, the opportunities for synergistic combination might not be discovered, and an opportunity to solve the new problem would be missed. For a number of well-known reasons, mentioned briefly below, researchers from different disciplines might indeed not be aware of one another's work.

Swanson coined the term "undiscovered public knowledge" to describe such situations. Portions of the solution to a new problem could already reside in the public literature, but their synergistic combination toward this new end might have never been considered. The suggestion alone re-invites researchers in all disciplines to consider the public literature of their own, and other, professional communities as a potential source of information that may be relevant to their own investigations. The suggestion challenges researchers in Information Retrieval (IR) to seek effective and efficient implementations of search mechanisms capable of detecting heretofore undetected relationships. To avoid redundant alerting of researchers, such implementations would need to be capable of screening out relationships between topics that are already present in the literatures of the topics.

Swanson labeled as "complementary but disjoint" topical literatures that are potentially combinable toward some new end, but that are not already bibliographically related [5]. Undiscovered public knowledge is related to the larger problem of "literature based discovery", wherein the body of research literature is employed as a source of knowledge by researchers. The literature would be seen to have status and information potential comparable to the laboratory. Bibliographic relations would include discussion of one another's concepts, or a common citation of specific papers [6]. If the literatures are not disjoint, the presumption is that some researchers have already addressed the particular combination of research topics that is in question. If the combination of research topics is already under investigation,

or has previously been investigated, then presumably the combination need not be redundantly brought to the attention of researchers [7] [8] [9].

Swanson [1] [10] [4] [11] [12], and Swanson and Smalheiser [13] demonstrated the existence of exemplary topical pairs of undiscovered public knowledge. Their search techniques combined repeated applications of world knowledge, domain expertise, and computerized tools to facilitate searches for words in the titles of articles in the MEDLINE dataset. The ARROWSMITH system [13] [14] [15] [16] was developed to facilitate sequences of searches involving words from article titles.

Even if the probability that a specific pair of research results might be synergistically combined is small, an extremely large number of such pairs are possible. But, to deal with large numbers of possible pairings, a way must be found to examine candidate literature pairings inexpensively. Some elementary statistical measures employed in [17] are shown here to be sufficient to be used in a supercomputing context for identification of promising pairings. Considering the number of potential pairs that might turn out to be exploitable, a very great amount of good might be done for society as a whole via better informed research communities.

Definitions

For purposes of this paper, a "research topic" or "topic" is a thematic area of research or scholarly activity. In this study, as in [1] and [17], Raynaud's Disease and dietary fish oil are the topics examined. A professional researcher may be involved with more than one topic. A research discipline may encompass multiple topics. A topic need not exactly coincide with a particular subject heading in any classification system.

The "topical literature", then, is the set union of all writings concerning the thematic area of the topic. In studies such as this one, there are practical limits on what documents, or portions of documents are conveniently reachable (e.g., what is stored in a machine readable form, or what is publicly available without some restrictive form of license). This study used the MEDLINE dataset and the associated MeSH classification system as both are freely available, and in electronic form. Note, however, that MEDLINE does not contain the whole text of articles, but only their titles, abstracts, and other bibliographic metadata. Thus, in this study, the topical literatures are approximated by the text of the MEDLINE article titles, and the abstracts of the articles, rather than the whole text of the articles themselves. In this study, like [1] and [17], topical literatures are defined relatively simplistically as those documents cataloged under a specific set of MeSH headings. 2,650 MEDLINE records so retrieved constituted the topical literature of Raynaud's Disease. 600 records similarly constituted the topical literature of dietary fish oil.

A "topical literature pair" or "pair" is a pair of topical literatures, paired simply for the purpose of considering the relationship between them.

Similarly, a "topical research community" or "community" is the set of all researchers professionally involved with the topic. As a practical matter, since written papers are the point of origination for all the information examined herein, the existence of those members of a community who are not authors of papers contributing to the topical literature is simply not detected in any of the analyses herein. As an underlying motivation in information retrieval research is for all members of a community to be better informed, readers as well as writers are included in what is termed a community.

Swanson further differentiated topical literatures as either "source" or "target", based on the direction taken by the human researcher while searching for a solution to a new research problem. As the analytical approach used herein is not specific to direction, both literatures in a pair are termed "candidate literatures" (i.e., candidates for designation as undiscovered public knowledge examples) while they are being examined for possible relationships between them.

Not all that is written is made "public". For example, commercial research laboratories do not distribute all their research results. Wanting to retain as much as possible of Swanson's terminology here, the designation "public" then should here be understood to mean "available for information retrieval study and for medical researcher use". If a similar study were conducted under the auspices of a commercial laboratory, they would presumably also include examination of proprietary documents concerning their research.

Roots of Incomplete Communication in Science

Public dissemination of research ideas, hypotheses, and results, with attendant debate and examination by peers, are fixtures of scientific research [18] [19] [20]. In order to begin looking for pairs of topical literatures that are complementary but disjoint, it is helpful to review factors that contribute to isolation of one researcher or group from another [21]. Factors impeding the communication process arise from a number of sources, including; researcher specialization, community normative processes and paradigms [19], limited professional contact with researchers outside one's immediate discipline and physical vicinity [22] [23] [24], tendencies to interact with known and trusted colleagues [25] [24], and the structure of reward systems (e.g., tenure, promotion, patent ownership, sole authorship) under which researchers are employed. Citation analysis has been used successfully to identify subgroups within a larger research community. Investigations of multidisciplinary teaming behavior [22] seem to bear out assumptions of a tendency toward specialization by researchers in that co-authoring and co-citations are relatively rare across disciplinary boundaries [23] [26]. Sandstrom [27] [28] compares the processes scientific researchers go through in selecting research areas to the processes subsistence foragers go through in selecting areas in which to forage for needed resources. Optimal foraging theory may aid understanding of how researchers choose to invest their resources, including their time, so as to acquire future resources. Sandstrom suggests [27] that the specific resource scientists may be said to be foraging for might be "information novelty", as opposed to information itself. That is, the information, or the conclusions supported by the information, must necessarily be in some way new in order to deserve a place in the research literature of the day, and thereby to indirectly sustain the researcher via the reward

systems. If such is the case, tools to identify topical pairings not heretofore examined (grazed) could be very desirable to researchers in many fields.

If research literature is oriented toward novelty, then the well-known would not be as well represented. As [1] [17] and this study use the MEDLINE research literature, there is the possibility that widely known medical information is simply not included in the collection. If so, incorporation of the researcher's own domain knowledge and world knowledge, as was done in [1] and [17], may prove indispensable. To attempt coverage of the more widely known information in an automated system such as this study considers, it may be necessary to also include survey, tutorial, or textbook literature.

How Much Undiscovered Public Knowledge Is There?

After initially publicizing the first sets of results from successful searches for undiscovered public knowledge, Swanson performed a more in-depth analysis of another example case, that of migraine and magnesium [4]. He enumerated a total of eleven intermediate topical literatures that might have given researchers clues to the roles of magnesium in migraine headache processes. The existence of multiple undetected relationships in this case, while this case is only a single data point, is disturbing. Missed applications of knowledge may turn out to be distressingly common.

Picking two topical literatures at random is very likely to result in a failure to find a complementary relationship. Even with the recent proliferation of supercomputers built of commodity processors operated in large clusters, computing costs are such that trying "all possible pairs" is unjustifiable.

The number of possible pairings is given by the binomial expansion coefficient. Given n objects, from which r are to be selected without replacement, and where the order of selection is unimportant, the number of such possible combinations is given by the formula

```
n!/((n-r)! * r!) .
```

For 10 topical literatures (n = 10), the number of possible topical literature pairs (viz., r = 2) is 45. For 100 topical literatures (n = 100), the number is 4,950. In the 2000 version of MeSH, there are 19,636 medical subject headings. If each subject heading is taken as definitive of a topical literature (viz., setting n = 19,636), 192,776,430 possible topical literature pairs exist to be analyzed.

Pair-wise examination of topics via supercomputing does not follow a line of reasoning that traces causality. Instead of considering inferences in a causal sequence (e.g., "A causes B, B causes C, therefore A causes C"), the supercomputing approach instead considers something more like "A is associated with B, C is associated with B, might A be associated with C?" The direction of causality may ultimately turn out to be that "C causes A". The computer does not consider either of the literatures C or A as either what Swanson terms "source" or "target", but treats both literatures simply as candidates to be considered. Relatedness between candidate pairs of topical literatures may be indicated

statistically, but determining the sequence of causality necessitates the application of some intelligence.

If ways can be found automatically, or in combination with researcher input, to efficiently search for topical literature pairs producing desirable results, the potential is there for the automated processing to contribute information of value to the research process. If computerized statistical screening is used over large numbers of topical areas already considered plausible by researchers, those pairs with the most promising statistics could be relatively inexpensively identified. A system implementation possibility would be to alert researchers to such pairings. Alternatively, for disciplines where the simple identification of plausible topical pairings is not an impediment to getting started, the analysis could be used in the process of prioritizing among research options.

Seeking Features Identifiable via Supercomputing

Prior work

Gordon & Lindsay performed similar work, to both validate Swanson's results in the case of Raynaud's disease and dietary fish oil, and to explore the possibility automated processing could be used to a greater degree [17]. They used readily computed statistical features (i.e., record frequency, token frequency, relative frequency, and token frequency divided by global frequency). Their analysis proceeded in the direction from a source literature of Raynaud's disease to a target literature of dietary fish oil. The strong, basically definitional links between Raynaud's disease and matters pertaining to blood flow mechanics [2] showed up in their work, Swanson's work, and the work described herein. However, their selection of blood flow terms involved the use of researcher consideration of the frequently occurring two-word terms related to blood flow. Other terms were also available, so perhaps blood flow might not have been the next topic the hypothetical researchers may have selected.

Calculation of the statistics used in [17] was replicated in this study, but in an implementation that is independent of the MEDLINE web sites. It was hoped some combination of these indicators would be useful in constructing a supercomputer-based approach to the examination of a very large number of candidate literature pairs.

Traditional search and research trajectories

Search path selection employs the world and domain knowledge of the searcher in selecting the paths felt to most probably yield desired results. The combined effects of this world and domain knowledge can be said to form an approximation function the researcher uses to estimate the desirability of a certain future problem state, and the path associated therewith. Foreknowledge of the ultimate outcome of selecting any particular path is usually only complete in the case of very small problem spaces.

In game theory, comparative analysis usually is performed, from the perspective of all players, out to some number of moves ahead [29]. Value appraisal among alternatives is done via some relatively

inexpensively computed form of approximation (e.g., using heuristics such as "positioning bishops on squares having long diagonals is good". In search and research, comparable heuristics can be exploited (e.g., "this author is always very succinct and informative", or "xyz will dissolve most hydrocarbons of comparable complexity"). Decisions concerning expenditure of resources (e.g., selecting and making a chess move, or performing the next step in a search or research sequence) can then be done on the basis of comparative desirability, however incompletely known.

In literature based discovery approaches, an implicit assumption is of a human searcher, searching for an unknown goal. The goal (target) literature will be recognized as valuable only after it is identified. If it were known to be of value at the outset, it could have simply been retrieved.

The research trajectory of beginning with one concept (e.g., a medical problem, or the availability of a new drug) and moving outward through a chain of reasoning involves considerable guidance from knowledgeable researchers. To scale to larger numbers of pairs being investigated, systemic dependencies on the available time of knowledgeable researchers should be minimized.

Supercomputing investigation trajectory

Involving human researchers at numerous points in the search for undiscovered public knowledge seemingly limits the scalability of the system as a whole to the capacity of the staff. Tools to support individual researchers are valuable, but would extend the progress only by improving an efficiency multiplier. Considering the expenses of research and the critical role research plays in extending science or a corporate product line, even a linear efficiency increase is a desirable thing. While a supercomputer-based filtering or alerting system may be the ultimate result of our work, we are hopeful the number of literature comparisons made possible will be substantial. It is, however, too early in the investigation to tell.

Rather than facilitating stages in a traditional scientific research or library search process, the supercomputer-as-filter approach assumes a given topical pair is related, without first proving this. Then, processor time is expended to explore the nature and strength of the relationship. Exploration proceeds simultaneously, for a large number of topical pairs, limited fundamentally by the number of processors that can be afforded.

In an extensively automated environment, numerical measures are necessary to provide some way to reliably identify relatedness. It is desirable that such functions be inexpensively calculable, and typically from data that is readily available in machine-readable form. This study particularly examined statistically based relatedness measures from [17], to determine if they appear viable for use in a supercomputing approach to the examination of very large numbers of possible topic pairs.

Relatedness measures so calculated for a given pair would then be compared, relatively, to those of other pairs. Any particularly outstanding values would then be identified to researchers. Researchers may then want to use ARROWSMITH, or other search and retrieval facilities, to pursue their investigations in

more depth.

Such a seemingly simplistic and wasteful approach is not practical in systems centered around human efforts, as the number of possible topical pairings will be huge. But, it is possible, reasonably economically, in a supercomputer environment, providing suitable calculable indications of relatedness can be found.

Emulating earlier studies, but using an expanded processing capacity

The system discussed herein was constructed for the examination of a very large number of topical literature pairs. As such, it could not be allowed to burden National Library of Medicine online resources. A complete copy of MEDLINE was therefore obtained for work on local computers. An investigation was then performed to attempt to emulate the statistical results of [17], after Swanson's originally announced pair of literatures [1] (i.e., Raynaud's disease and dietary fish oil), under this different implementation architecture.

The following four statistical measures from [17] were computed, but using Single-Instruction Multiple Datastream (SIMD) parallel processing. All these steps first involved the selection of the records constituting the topical literature, and subsequent parsing, stoplisting, and stemming of words from those records. Unlike ARROWSMITH and the studies by Swanson that used only text from titles, this study and [17] also used the text of abstracts and MeSH headings. Terminology used below attempts to follow [17].

(1) Token Frequency: The number of times a certain token occurs within the topical literature. Note that multiple occurrences of a token within a single record were counted as separate occurrences.

To determine token frequency in this study, each word surviving stoplisting was written to a token file. Each pair of consecutive words surviving stoplisting was also written to the token file. Tokens here, and in [17], were thus one word, or two word phrases. The token file was then alphabetically sorted, and a token frequency file formed, including both the tokens and the count of their use within the topical literature.

An array of tokens and occurrence counters was thereby produced. The array was truncated to only include the set of all tokens from the topical literature that occurred more than once. Elimination of tokens occurring only once reduces this dataset size substantially (e.g., 63% of the terms from the "Raynaud's disease" literature examined occurred only once, as did 60% of the terms associated with the corresponding "fish oil" literature).

(2) Record Frequency: The number of records in a topical literature containing at least one occurrence of a particular token.

To determine record frequency in this study, a document counter field associated with each token retained in the token frequency processing was incremented, upon completion of processing of each document, for the tokens encountered in that document.

(3) Relative Frequency: Defined numerically as the number of documents within the topical literature that contain at least one occurrence of the token sought, divided by the number of documents within all of MEDLINE that contain at least one occurrence of the token sought.

To determine relative frequency in this study, the number of documents within all of MEDLINE that contain at least one occurrence of the token sought was calculated in an analogous manner to that of record frequency. Every record in MEDLINE was parsed and processed, as had been done for records of the topical literature. A global frequency counter field associated with each token retained in the token frequency processing was incremented, upon completion of processing of each document, for the tokens encountered in that document. Having used the token frequency counter table as a base, tokens encountered here that were not retained by the previous token frequency processing (i.e., those tokens not occurring more than once in the topical literature subset of all of MEDLINE) were ignored.

(4) tf*igf ("token frequency times inverse global frequency"), defined mathematically as

> (token frequency)*log((number of records in MEDLINE)/(number of records in MEDLINE containing the token sought)).

In this study, the number of records in MEDLINE, in print as of the date examined, was a constant 5,952,084.

Implementation notes

Selection by date of publication

MeSH headings have dates of applicability, and can subdivide, merge, or be relabeled over time. To look at all, or large portions of the chronology of MEDLINE, compensation for MeSH changes is necessary. As this study examined records with dates of publication spanning twenty-four years, program code was developed to select topical literature records based on a combination of both the MeSH headings used to index the article, and the date of publication of the article.

The program code supports the use of potentially different MeSH headings, presumably concerning the same topic, over time. For example, due to MeSH changes, the topical literature concerning dietary fish oil is defined using the MeSH term "Fish Liver Oils" for dates of publication from 1963 to 1966, and the term "Fish Oils" for dates of publication subsequent to 1966. The topical literature for Raynaud's syndrome, Raynaud's disease, and Raynaud's phenomenon is defined using the MeSH term "Raynaud's Disease", for all dates of publication.

Examining only the pre-discovery literature

This study used only records of the topical literatures with document dates of publication before January 1, 1989. As explained in [30], while the date of Swanson's first published claim of relatedness between Raynaud's disease and dietary fish oil [1] was 1986, [3], published in 1989, was the first medical confirmation of the relationship published.

Although [30] found no citations of Swanson's various publications in medical literature by other authors, including [3], medical publications subsequent to [3] might have been influenced by the content of [3], and were therefore held out of this analysis. A subsequent analysis, not discussed herein, showed the medical confirmation of this relationship did not substantively change the nature of the vocabulary being used in either of the post-confirmation topical literatures.

Stoplisting and stemming

The stoplist from ARROWSMITH was used as a baseline in this study, in the interest of facilitating comparison of results. The original ARROWSMITH stoplist included 7,997 non-numerical terms. No two-word phrases were included. The parser used in this study algorithmically eliminates all-numeric entities, thus obviating 248 numerical entries from the ARROWSMITH stoplist.

A total of 1,349 adaptations were then made to the ARROWSMITH list, on a per-topic basis, through examination of the most frequently used terms that survived the Swanson stoplist. This filtering step approximates ARROWSMITH-based involvement of researchers to screen the output terms for applicability [13]. And, the manual removal of obvious discards from the summary tables in [17]. However, better automated techniques will be needed to perform this step if undiscovered public knowledge searches are to be conducted in bulk quantities.

ARROWSMITH includes algorithmic facilities to handle some forms of stemming, particularly for plurals. These mechanisms are separate from the stoplist provided. Facilities were added to the system discussed herein to accomplish stemming. Stemming is done for explicitly listed words, in a case-by-case manner, for those most frequently used words surviving conventional stoplisting.

Statistics of the MEDLINE dataset used

A total of 13,037,195,962 bytes of data are included in the 2000 version of MEDLINE used herein. This dataset includes 10,171,686 titles, and averages 1,282 bytes of data per MEDLINE record. 107,315,659 MeSHs were used, giving an average of 10.6 terms used per MEDLINE record. 5,149,362 abstracts are included. Thus, 50.6% of the records have an abstract.

By contrast, the CANIS laboratory encountered 9,619,223,757 Bytes total dataset size in their work using MEDLINE in 1996 [31], reflecting a dataset volumetric growth of 36% in roughly 4 years time.

Results Achieved

Unlike [17], which proceeded from a statistical evaluation of the literature of Raynaud's disease to an examination of the literature concerned with blood flow, and thereby on to dietary fish oil, this study examined links originating from either the topical literatures of Raynaud's disease or dietary fish oil. All interrelationships found are reported herein, as contrasted with [17] that removed links not plausibly "terminals", that is, not seemingly an ameliorative agent against Raynaud's disease. The tables of results herein are correspondingly more lengthy than those of [17]. Like [17], results are provided for both single-word and two-word (phrase) terms.

Tokens related to blood flow were found in the sets of results from both Raynaud's disease and dietary fish oil. As such, a filter based around the mechanism described herein should be able to identify that some form of relation exists between Raynaud's disease and dietary fish oil, via some terms and phrases associated with blood flow. The system herein has no notion of causality or directionality of the linkage, and relationships found may not concern corrective or ameliorative effects presumably sought.

Table 1 lists the quantities of tokens processed for the topical literatures and for MEDLINE overall. Table 2 lists single-word tokens found in the literature of Raynaud's disease, ranked in order of the relatedness measure named at the head of each column. Table 3 is similar, but for two-word tokens. Tables 2 and 3 largely duplicate the results of [17] in establishing a statistical link from Raynaud's disease to blood flow.

Statistic	Raynaud's Disease	Dietary Fish Oil	MEDLINE Overall
Number of documents in the literature	2,658	660	5,952,084
Number of tokens	8,659	4,949	5,593,396
Number of tokens that are word pairs	2,861 33%	1,885 38%	3,842,023 69%
Number of tokens that occurred only once and were subsequently ignored	5,454 63%	2,979 60%	3,776,833 68%

Table 1 Dataset quantities and token statistics for the topical literatures, and for MEDLINE overall.

Table 4 lists single-word tokens found in the literature of dietary fish oil, ranked in order of the

relatedness measure named at the head of each column. Table 5 is similar to table 4, but for two-word tokens.

Structurally, all these tables are patterned after [17]. But, tables 4 and 5 differ significantly in their starting point, the topical literature of dietary fish oil. These tables show an attempt to "work backward", going toward the disease instead of beginning there and moving away. Some use of tokens related to the mechanics of blood flow were indeed found. The single-word token rankings of table 4 show some blood-related tokens (e.g., blood, platelet) that can be considered applicable to a discussion of Raynaud's disease. Table 4 also contains some blood-related tokens that pertain much more broadly to the circulatory system (i.e., heart, artery, coronary), and which are probably not generally associated with a discussion of the effects of Raynaud's disease in the extremities.

Further, relationships between topical literatures based on the single word "blood" are not particularly satisfying in that blood is the word that occurs most, of all words surviving the stoplist. Taken over the time period of this study (i.e., with dates of publication prior to 1989), the word blood occurs 958,702 times. A great many things may be said to be somehow related to blood.

The results in table 5 are more to the point, listing blood platelet, blood viscosity, blood coagulation, and platelet membrane among the more high-ranking terms.

Rank	Token Frequency	Record Frequency	Token Frequency * Inverse Global Frequency	Relative Frequency
1	Raynaud's	Raynaud's	Raynaud's	antivibration
2	blood	scleroderma	scleroderma	crst-syndrome
3	scleroderma	finger	finger	rock-drill
4	finger	blood	vibration	Raynaud's
5	vibration	vasculature	vasculature	aca-negative
6	vasculature	skin	blood	Crest'
7	artery	artery	skin	acrosyndromes
8	skin	vibration	artery	c.r.s.t
9	capillary	sclerosis	vasospasm	crst
10	sclerosis	lupus	capillary	cold-provoked
11	lupus	erythematosus	nifedipine	cryopathies
12	rheumatic	rheumatic	sclerosis	dosa

13	erythematosus	vasospasm	lupus	endarteriosis
14	nifedipine	arteriosclerosis	erythematosus	microvasculotissular
15	pulmonary	arthritis	rheumatic	non-vwf
16	vasospasm	calcinosis	obliterans	postreserpine
17	hypertension	vasodilator	cool	Reynaud's
18	esophagus	capillary	antinuclear	taylor-pelmear
19	arthritis	obliterans	vasodilator	Thibierge-weissenbach's
20	vasodilator	ischemia	calcinosis	vasoneuroses
21	calcinosis	hypertension	esophagus	hand-transmitted
22	obliterans	collagen	arthritis	aca-positive
23	platelet	arm	ketanserin	hexopal
24	cool	muscle	thermography	acrosyndrome
25	antinuclear	antinuclear	hypertension	arteriospastic
26	ischemia	telangiectasis	pulmonary	chipping-hammer
27	arteriosclerosis	esophagus	pss	fsbp
28	muscle	pulmonary	vwf	tetranicotinoylfructose
29	collagen	ulcer	telangiectasis	sclerodactyly
30	viscosity	thromboangiitis	arteriosclerosis	morpho-oscillography
31	ulcer	hands	viscosity	hand-arm
32	thermography	cool	rp	anticentriole
33	thorax	thorax	ischemia	vasoneurosis
34	systolic	heart	biofeedback	vasopastic
35	arm	vasoconstrictor	sle	chainsaw
36	ketanserin	neoplasm	crest	riveters
37	neoplasm	nifedipine	thromboangiitis	sclerodactylia
38	heart	vasomotor	platelet	acrosclerosis

39	sle	thermography	hands	acrocyanosis
40	biofeedback	circumscribed	collagen	anticentromere
41	hands	viscosity	anticentromere	acrorhigosis
42	vasoconstrictor	systolic	systolic	bradilan
43	crest	kidney	arm	cenp-a
44	telangiectasis	microcirculation	cryoglobulins	cenp-b
45	rp	crest	Sjogren's	chippers
46	pss	protein	vasoconstrictor	cortisolaemia
47	vwf	arteritis	crst	kpu
48	prostaglandin	nail	thorax	multifinger
49	bone	cryoglobulins	nailfold	multispecialist
50	thromboangiitis	gangrene	ulcer	panergon
51	Sjogren's	adrenergic	vinyl	pipework
52	microcirculation	platelet	nail	quarriers

Table 2. Analysis of single-word tokens contained in Raynaud documents published prior to medical confirmation of Swanson's assertion. Note that tokens occurring in only a very small number of documents in MEDLINE cause extremely large values of Relative Frequency, thus only those terms occurring in at least 3 MEDLINE documents were used in the tf*igf and the Relative Frequency ranking shown here.

Rank	Token Frequency	Record Frequency	Token Frequency * Inverse Global Frequency	Relative Frequency
1	lupus erythematosus	lupus erythematosus	lupus erythematosus	bush cutter
2	blood viscosity	thromboangiitis obliterans	thromboangiitis obliterans	cold-induced Raynaud's

3	thromboangiitis obliterans	arthritis rheumatic	blood viscosity	dead finger
4	arthritis rheumatic	arteriosclerosis obliterans	finger systolic	pulmonary Raynaud's
5	arteriosclerosis obliterans	scleroderma circumscribed	arteriosclerosis obliterans	Raynaud's vasospasm
6	scleroderma circumscribed	blood viscosity	scleroderma circumscribed	vasospasm Raynaud's
7	rheumatic arthritis	rheumatic arthritis	nailfold capillary	vibration Raynaud's
8	vinyl chloride	hypertension pulmonary	vinyl chloride	scleroderma Raynaud's
9	pulmonary hypertension	finger systolic	arthritis rheumatic	acral ischemia
10	finger systolic	angina pectoris	pulmonary hypertension	cold-induced vasospasm
11	nailfold capillary	blood protein	finger blood	finger systolic
12	angina pectoris	erythematosus sle	carpal tunnel	chainsaw vibration
13	carpal tunnel	systolic blood	rheumatic arthritis	cutis Raynaud's
14	systolic blood	nailfold capillary	hand-arm vibration	finger blanching
15	finger blood	adrenergic beta-antagonists	finger vwf	finger rewarming
16	hypertension pulmonary	pulmonary hypertension	calcinosis Raynaud's	intraarterial reserpine
17	blood protein	vinyl chloride	angina pectoris	pedestal grind
18	pulmonary fibrosis	brachial artery	systolic blood	spasm Raynaud's
19	myocardial infarction	carpal tunnel	sclerosis pss	tip ulcer
20	subclavian artery	diabetes angiopathy	prostaglandin e	vasculature acrosyndromes
21	brachial artery	finger blood	hypertension pulmonary	vibration vasculature

22	erythematosus sle	calcinosis Raynaud's	finger skin	finger cool
23	hand-arm vibration	sclerosis pss	subclavian artery	ena ribonuclease
24	adrenergic beta-antagonists	blood platelet	sclerosis scleroderma	finger hemodynamics
25	diabetes angiopathy	finger vwf	finger cool	hand-held vibratory
26	diabetes mellitus	liver cirrhosis	erythematosus sle	headache Raynaud's
27	prostaglandin e1	pulmonary fibrosis	brachial artery	numb finger
28	sclerosis pss	diabetes mellitus	pulmonary fibrosis	pneumatic percussion
29	biliary cirrhosis	sclerosis scleroderma	biliary cirrhosis	purpura Raynaud's
30	blood platelet	skin ulcer	ulnar artery	riveting hammer
31	liver cirrhosis	subclavian artery	prostaglandin e1	stone cutters
32	calcinosis Raynaud's	biliary cirrhosis	diabetes angiopathy	vinblastine-bleomycin chemotherapy
33	finger vwf	hand-arm vibration	stellate ganglia	finger vwf
34	prostaglandin e	myocardial infarction	skin ulcer	cool finger
35	skin ulcer	blood coagulation	blood protein	bush cleaners
36	blood coagulation	cirrhosis biliary	adrenergic beta-antagonists	chipping hammer
37	sclerosis scleroderma	testes neoplasm	vasculature acrosyndromes	finger ulcer
38	finger skin	polyarteritis nodosa	chipping hammer	hand-transmitted vibration
39	stellate ganglia	anemia hemolytic	polyarteritis nodosa	scleroderma gs
40	polyarteritis nodosa	finger cool	cirrhosis biliary	hand-arm vibration
41	ulnar artery	finger skin	myocardial infarction	anticentromere antibody-positive

42	cirrhosis biliary	prostaglandin e1	esophagus dysmotility	calcinosis Raynaud's
43	finger cool	adrenergic alpha-antagonists	fingertip blood	finger photoplethysmography
44	testes neoplasm	arteriovenous fistula	liver cirrhosis	finger pressures
45	anemia hemolytic	artery blood	capillary blood	humeral arteriographic
46	pulmonary vasculature	red blood	thromboxane synthetase	ischemia Raynaud's
47	capillary blood	stellate ganglia	blood platelet	nicotinate hexopal
48	pulmonary artery	erythematosus discoid	spasm Raynaud's	numb paresthesia
49	thromboxane a2	varicose vein	testes neoplasm	oxalate praxilene
50	red blood	cervical vertebrae	pulmonary vasculature	periflux blood
51	adrenergic alpha-antagonists	esophagus dysmotility	thromboxane a2	rock drillers
52	artery blood	ganglia autonomic	diabetes mellitus	saw vibration
53	cervical vertebrae	polyvinyl chloride	ischemic necrosis	sclerodactyly telangiectasias
54	polyvinyl chloride	pulmonary vasculature	polyvinyl chloride	thorax gangliectomy
55	thromboxane synthetase	ulnar artery	sclerosis ss	thorax sympathectomies

Table 3. Analysis of two-word tokens contained in Raynaud documents published prior to medical confirmation of Swanson's assertion. Note that tokens occurring in only a very small number of documents in MEDLINE cause extremely large values of Relative Frequency, thus only those terms occurring in at least 3 MEDLINE documents were used in the tf*igf and the Relative Frequency ranking shown here.

Rank	Token Frequency	Record Frequency	Token Frequency * Inverse Global Frequency	Relative Frequency
1	oil	fish	oil	fo-fed
2	fish	oil	fish	fish-oil
3	lipid	lipid	lipid	max-epa
4	platelet	liver	platelet	cod-liver
5	liver	blood	cholesterol	mo-fed
6	cholesterol	cholesterol	n-3	partially-hydrogenated
7	phospholipid	platelet	cod	maxepa
8	blood	triglyceride	eicosapentaenoic	oil-enriched
9	triglyceride	eicosapentaenoic	phospholipid	7-methyl-7-hexadecenoic
10	vitamin	cod	liver	Epa'
11	lipoprotein	phospholipid	triglyceride	epa-fed
12	n-3	polyunsaturated	epa	phho
13	cod	vitamin	omega-3	vegetable-oil
14	eicosapentaenoic	arachidonic	lipoprotein	menhaden
15	membrane	lipoprotein	menhaden	omega-3-fatty
16	arachidonic	n-3	vitamin	14ceicosapentaenoic
17	epa	coronary	polyunsaturated	3-fatty
18	polyunsaturated	prostaglandin	n-6	hoplostethus
19	prostaglandin	membrane	cod-liver	mrl-mp-lpr
20	coronary	heart	arachidonic	txa3
21	corn	17-eicosapentaenoic	thromboxane	cetoleic
22	e	corn	corn	phfo
23	thromboxane	thromboxane	dha	omega-3

24	heart	docosahexaenoic	docosahexaenoic	3hdihydroalprenolol-binding
25	omega-3	epa	linoleic	bu2sncl2
26	menhaden	omega-3	hydrogenated	eicosapentanoic
27	linoleic	linoleic	e	epa-enriched
28	ldl	hdl	ldl	hho
29	n-6	menhaden	prostaglandin	l-485
30	hdl	n-6	blood	lco
31	diabetes	atherosclerosis	17-eicosapentaenoic	lna-rich
32	dha	artery	hdl	marine-oil
33	artery	b2	maxepa	mepyramine-treated
34	atherosclerosis	ldl	leukotriene	phfo-fed
35	docosahexaenoic	muscle	coronary	po-fed
36	cod-liver	bleeding	membrane	pusfa
37	microsomes	cardiovascular	pufa	ra-supplemented
38	leukotriene	hydrogenated	safflower	sunflowerseed-oil
39	hydrogenated	vasculature	atherosclerosis	tfo
40	arthritis	cod-liver	mo	u-14c-proline

Table 4. Analysis of single tokens contained in Fish Oil documents published prior to medical confirmation of Swanson's assertion. Note that tokens occurring in only a very small number of documents in MEDLINE cause extremely large values of Relative Frequency, thus only those terms occurring in at least 3 MEDLINE documents were used in the tf*igf and the Relative Frequency ranking shown here.

Rank	Token Frequency	Record Frequency	Token Frequency * Inverse Global Frequency	Relative Frequency

1	fish oil	fish oil	fish oil	docosahexaenoic dha
2	liver oil	liver oil	liver oil	fish oil-enriched
3	cod liver	cod liver	cod liver	fish oil-treated
4	corn oil	blood platelet	menhaden oil	oil hho
5	vitamin e	corn oil	cod-liver oil	oil menhaden
6	menhaden oil	menhaden oil	vitamin e	fish oil
7	cod-liver oil	thromboxane b2	corn oil	lard corn
8	thromboxane b2	hdl cholesterol	safflower oil	oil fo
9	blood platelet	lipoprotein hdl	thromboxane b2	cod-liver oil
10	safflower oil	cod-liver oil	hdl cholesterol	fish oil-fed
11	hdl cholesterol	lipoprotein ldl	leukotriene b4	fish oil-supplemented
12	leukotriene b4	safflower oil	hydrogenated fish	oxidized fish
13	lipoprotein hdl	leukotriene b4	olive oil	fish oil-induced
14	lipoprotein ldl	n-3 polyunsaturated	n-3 polyunsaturated	hardened fish
15	olive oil	diabetes mellitus	blood platelet	hydrogenated peanut
16	membrane phospholipid	membrane lipid	coconut oil	lipid cardiac
17	diabetes mellitus	coconut oil	linseed oil	lipid haemostatic
18	coronary artery	lipoprotein vldl	membrane phospholipid	lipid selenium
19	coconut oil	arthritis rheumatic	herring oil	ltb4 ltb5
20	membrane fluidity	membrane phospholipid	lipoprotein hdl	maxepa fish
21	linseed oil	coronary heart	mackerel oil	menhaden oil-fed
22	n-3 polyunsaturated	ldl cholesterol	beef tallow	ocean sunfish
23	hydrogenated fish	microsomes liver	lipoprotein ldl	oil hpo
24	ldl cholesterol	coronary artery	membrane fluidity	oil tfo
25	blood viscosity	linseed oil	fish liver	redfish sebastes

26	erythrocyte membrane	olive oil	menhaden fish	undistended jugular
27	lipoprotein vldl	thromboxane a2	ldl cholesterol	oil maxepa
28	membrane lipid	herring oil	vegetable oil	oil mo
29	thromboxane a2	oil fo	intimal hyperplasia	menhaden oil
30	herring oil	vegetable oil	fish oil-fed	oil fish
31	vegetable oil	fish liver	omega-3 polyunsaturated	liver oil
32	beef tallow	menhaden fish	primrose oil	menhaden fish
33	fish liver	oil co	thromboxane a2	oil clo
34	intimal hyperplasia	oil mo	oil fo	cod liver
35	lipoprotein cholesterol	6-ketoprostaglandin f1	prostaglandin e	epa fish
36	arthritis rheumatic	beef tallow	blood viscosity	ethoxycoumarin o-dealkylase
37	coronary heart	blood coagulation	oil mo	hoplostethus atlanticus
38	lipid peroxidation	erythrocyte membrane	lipoprotein vldl	mackerel oil
39	liver microsomes	membrane fluidity	n-3 pufa	mola mola
40	mackerel oil	rheumatic arthritis	soy oil	n-3 Pufa's
41	menhaden fish	blood glucose	coronary artery	oil oo
42	microsomes liver	blood lipid	peroxisomal beta-oxidation	oil phho
43	rheumatic arthritis	blood viscosity	membrane lipid	polyunsaturated fish
44	soy oil	hydrogenated fish	lipoprotein cholesterol	polyunsaturated n-3
45	primrose oil	low-density lipoprotein	oil co	sebastes marinus
46	vitamin d3	platelet phospholipid	erythrocyte membrane	sunfish mola

47	fish oil-fed	primrose oil	diabetes mellitus	trienoic prostaglandin
48	low-density lipoprotein	hydrogenated coconut	platelet phospholipid	tuna liver
49	omega-3 polyunsaturated	phosphatidylcholine pc	n-6 pufa	herring oil
50	peroxisomal beta-oxidation	lipid peroxidation	salmon oil	eicosapentaenoic epa
51	platelet membrane	lipoprotein cholesterol	vitamin d3	hydrogenated fish
52	prostaglandin e	mellitus insulin	lipid peroxidation	n-3 pufa
53	blood coagulation	n-3 pufa	coronary heart	cod oil
54	oil fo	omega-3 polyunsaturated	oxidized fish	hydrogenated beef
55	platelet phospholipid	platelet membrane	platelet membrane	mrl-mp-lpr lpr
56	red blood	prostaglandin e2	low-density lipoprotein	n-6 pufa

Table 5. Analysis of two-word tokens contained in Fish Oil documents published prior to medical confirmation of Swanson's assertion. Note that tokens occurring in only a very small number of documents in MEDLINE cause extremely large values of Relative Frequency, thus only those terms occurring in at least 3 MEDLINE documents were used in the tf*igf and the Relative Frequency rankings shown here.

Additional notes on the relationships found

Seeking statistical relationships between the two candidate literatures is effective, regardless of the literature chosen as the source literature (or problem description, or starting point). Table 6 lists the tokens that simultaneously occur in the 100 highest-ranked tokens from both literatures.

Token	Rank within Raynaud's Disease Literature	Rank within fish oil literature
Token Frequency		

diabetes mellitus	26	17
erythrocyte deformity	71	86
vasculature reactivity	69	68
rheumatic arthritis	7	43
thromboxane a2	49	29
coronary artery	70	18
myocardial infarction	19	66
prostaglandin e	34	52
arthritis rheumatic	4	36
vitamin e	100	5
thromboxane b2	75	8
blood viscosity	2	25
lupus erythematosus	1	97
red blood	50	56
blood coagulation	36	53
blood platelet	30	9
angina pectoris	12	75
Record Frequency		
diabetes mellitus	28	15
rheumatic arthritis	7	40
thromboxane a2	68	27
6-ketoprostaglandin f1	78	35
myocardial infarction	34	100
arthritis rheumatic	3	19
blood viscosity	6	43
red blood	46	72

blood coagulation	35	37
blood platelet	24	4
angina pectoris	10	80
(Token Frequency) * (Inverse Global Frequency)		
diabetes mellitus	52	47
erythrocyte deformity	83	98
vasculature reactivity	68	67
rheumatic arthritis	13	66
thromboxane a2	51	33
prostaglandin e	20	35
arthritis rheumatic	9	69
vitamin e	99	6
blood viscosity	3	36
red blood	72	82
blood coagulation	57	92
blood platelet	47	15
Relative Frequency		
No tokens in common from the 100 highest-ranked.		

Table 6. Two-word tokens occurring in the 100 highest-ranked tokens for both the Raynaud's disease and dietary fish oil literatures, using four statistical measures.

Like [17], and unlike Swanson's results, no simultaneously high rankings were found here concerning red blood cell rigidity. Unlike [17], no uses of the phrase "digital artery" occurred here because the ARROWSMITH-based stoplist used suppresses the word "digital". As, by convention in [17], the two-word phrases are built from two consecutive words that each survive the stoplist, the phrase "digital artery" can therefore not occur in this study.

Like [17], the ranking results of the relative frequency statistic exhibits little resemblance to the results of the other three statistics. Per the formula definition of relative frequency, if only one document, which

also happens to be a member of the topical literature of interest, contains a given token, that token will have a relative frequency of 1.0. This unnaturally high value falls off rapidly as the denominator, the number of documents in all of MEDLINE containing that token, increases (e.g., two documents containing that terms, only one of which is a member of the literature in question, reduces the relative frequency to 0.5).

This study uses a cutoff, requiring that a token occur in at least three documents in all of MEDLINE. But, the cutoff level is arbitrary, and it may be that obscure (undiscovered) simultaneous use of a common token may only occur a very small number of times. The numerical effects of the definition and the cutoff level imposed are viewed as substantial sources of perturbation for very-low-frequency terms. Table 7 shows the considerable variation in ranking by the relative frequency measure that results within the Raynaud's disease topical literature by simply varying low values of the token occurrence cutoff level.

No cutoff any occurrence in MEDLINE	**2 or more occurrences in MEDLINE**	**3 or more occurrences in MEDLINE**	**4 or more occurrences in MEDLINE**
10-minute-immersion	acro-syndromes	antivibration	antivibration
acro-syndromes	antivibration	crst-syndrome	rock-drill
aibut	crocq	rock-drill	Raynaud's
anti-m-a	crst-syndrome	Raynaud's	aca-negative
anti-rna-p	digit-to-brachial	aca-negative	Crest'
antivibration	exposure-to-cold	Crest'	acrosyndromes
chipping-hammers	finger-cooling	acrosyndromes	c.r.s.t
cold-pressor-induced	frequency-weighted	c.r.s.t	crst
crest-associated	hand-tool	crst	taylor-pelmear
crocq	Raynaud'	cold-provoked	hand-transmitted
crst-syndrome	resympathectomy	cryopathies	aca-positive
cryodynamic	rock-drill	dosa	hexopal
Ctd'	scalenus-anticus	endarteriosis	acrosyndrome
c-water	sd-pattern	microvasculotissular	arteriospastic

defibrinogating	white-finger	non-vwf	chipping-hammer
digit-to-brachial	Raynaud's	postreserpine	fsbp
dpit	aca-negative	Reynaud's	tetranicotinoylfructose
exposure-to-cold	Crest'	taylor-pelmear	sclerodactyly
finger-cooling	acrosyndromes	Thibierge-weissenbach's	morpho-oscillography
frequency-weighted	c.r.s.t	vasoneuroses	hand-arm
fvop	crst	hand-transmitted	anticentriole
handsyndromes	cold-provoked	aca-positive	cenp-c
hand-tool	cryopathies	hexopal	cl115
homoiothermics	dosa	acrosyndrome	vasoneurosis
humero-digital	endarteriosis	arteriospastic	vasopastic
ik-h	microvasculotissular	chipping-hammer	chainsaw
middlewest	non-vwf	fsbp	riveters
n.c.m	postreserpine	tetranicotinoylfructose	sclerodactylia
na-flu	Reynaud's	sclerodactyly	acrosclerosis
phendilin	taylor-pelmear	morpho-oscillography	acrocyanosis
presclerodermal	Thibierge-weissenbach's	hand-arm	anticentromere
pseudo-scleroderma	vasoneuroses	anticentriole	bradilan
pss-overlap	hand-transmitted	cenp-c	cenp-a
purpura-arthralgia	aca-positive	cl115	cenp-b
Raynaud'	hexopal	vasoneurosis	chippers
resympathectomy	acrosyndrome	vasopastic	multispecialist
r'group	arteriospastic	chainsaw	quinine-urea
roadbreakers	c.r.e.s.t	riveters	fsp
rock-drill	caulkers	sclerodactylia	chloride-associated

s.u.s.h	chipping-hammer	acrosclerosis	ketanserin's
scalenotomia	cineesophagram	acrocyanosis	megacapillaries
scalenus-anticus	cold-test	anticentromere	mortuus
sd-pattern	dehydronifedipine	acrorhigosis	pletysmography
ss-rp	dehydronifedipinic	bradilan	lumberjack
tns-dependent	f.s.p	cenp-a	a2r
unguis-like	foot-warming	cenp-b	barbotan
vwfd	fsbp	chippers	depth-sense
white-finger	m.e.c	cortisolaemia	fc-gamma-receptor
Raynaud's	omeral	kpu	fellers
aca-negative	poikilo-	multifinger	hexanicit

Table 7. Effect of varying document occurrence cutoff on the relative frequency ranking of tokens from the Raynaud's disease literature.

Conclusions

Statistical examination of literature pairs can produce some form of linkages through an intermediate literature, even in the absence of domain expertise, and independent of any direction of causality which may exist. While this is far short of a generalized knowledge acquisition mechanism, such statistical information could be useful in informational way to the process of research prioritization. Without significant expenditure of researcher time, and using the relatively inexpensive supercomputer architecture based around clusters of commodity computers, statistical clues can be acquired that could be part of the prioritization process in research.

Projecting slightly ahead, to offices equipped with more powerful workstations, or with network-of-workstation architectures to cluster computer assets, similar mechanisms could be used to rapidly identify points of similarity in other literatures. Application areas might include background investigation work during the writing of a news article or contract proposal, or in laying out one's thoughts in the outline of a paper. The process of launching such investigations could be built into word processors or other tools. It would not be difficult to recast the software developed herein as something launched by, say, an emacs macro, when an author right-clicks on a term in a document that is also a MeSH term.

Future Work

Initial trials were performed with a noun phrase parser from MEDLINE work of the CANIS laboratory and the University of Arizona [31]. Their work used the parser configured to produce noun phrases of up to seven words in length. However, when run on the MEDLINE dataset, 86% of the unique phrases detected occurred only once. As this appeared to provide very little basis for statistical evidence, this project worked with 1- and 2-word terms, as in [17]. Additional work appears needed in this area, considering the relatively desirable results achieved by the two-word phrases. Some combination of algorithm selection and parameter tuning may be valuable in providing more context-rich phrases for analysis.

In another alternative implementation, a minimalist stoplist approach should be tried. It may be that the statistics being run, on the very many more noun phrases that result, will satisfactorily deal with the problem of de-emphasizing less important phrases.

It is planned to incorporate the UMLS Metathesaurus [32], to deal with large numbers of vocabulary problems arising from different canonical forms, use of synonyms, and "is a", "part of", and other structural relationships.

To build a more extensive automatic test facility, citation analysis should be explored [6]. Some form of automatic identification of "discovered public knowledge" would need to be incorporated into any operational facility built along these lines.

Open Questions

Are discoveries found in this way undiscovered public knowledge, or perhaps just a reification of "common knowledge"? The scope of the MEDLINE collection is "research literature", and "1963-present". As such, older materials are not restated therein (e.g., Raynaud's original paper [2], textbooks, or concepts that now are part of the assumed medical education process), although some form of summarized citation may exist in the whole text. Inferring something from MEDLINE via extensive statistical manipulation such as was done in this study may seemingly yield an item of "common knowledge", and therefore be uninformative. For example, Raynaud's 1862 paper discussed processes related to blood flow in the extremities, so automatic regeneration of the relationship between Raynaud's disease and blood flow could be said to be entirely anticlimactic.

Acknowledgements

The author wishes to thank the other members of the team engaged in investigations and proposals related to undiscovered public knowledge at the Graduate School of Library and Information Science at

the University of Illinois (F. Wilfrid Lancaster, Carole Palmer, Mark Spasser, and David Dubin) [33], Professor Bruce R. Schatz, Richard Berlin, M.D., and Neil Smalheiser, M.D., Ph. D., for their many points of discussion and insights.

References

[1] Swanson, Don R. "Undiscovered Public Knowledge." *Library Quarterly* 56 (April 1986):103-18.
[2] Raynaud, Maurice. De l'Asphyxie Locale et de la Gangrène Symétrique des Extrémités. L. Leclerc, Paris. 1862.
[3] DiGiacomo, R.A., Kremer, J.M., & Shah, D. H. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. American Journal of Medicine, 86(2), 158-164, 1989.
[4] Swanson, Don R. "Migraine and Magnesium: Eleven Neglected Connections." Perspectives in Biology and Medicine, 31(4), 526-557. 1988.
[5] Swanson, Don R. "Complementary Structures in Disjoint Science Literatures". In A. Bookstein, Y. Chiarmella, G. Salton, & V.V. Raghavan (Editors) Proceedings of the Fourteenth Annual International ACM SIGIR Conference, October 13-16, 1991, pp. 280-289. ACM Press, Chicago, IL.
[6] Swanson, Don R.. "The Absence of Co-Citation as a Clue to Undiscovered Causal Connections". In Scholarly Communication and Bibliometrics, Christine L. Borgman, editor. Sage, Newbury Park, CA. 1990.
[7] Martyn, J. Unintentional duplication of research. New Scientist, 377, 1964, 338.
[8] Martyn, J. Literature Searching Habits and Attitudes of Research Scientists. Boston Spa, British Library, 1987.
[9] Floren-Romero, Maria Soledad 1994. *The Impact of Information Loss on Research: A Case Study in the Dominican Republic*. Ph.D. Thesis, UIUC GS-LIS.
[10] Swanson, Don R. "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge." Perspectives in Biology and Medicine, 30, 1, pp. 7-18. Autumn 1986.
[11] Swanson, Don R. "Two Medical Literatures that are Logically but not Bibliographically Connected." Journal of the American Society for Information Science, 38(4): 228-233, 1987.
[12] Swanson, Don R. "Medical Literature as a Potential Source of New Knowledge." Bulletin of the Medical Library Association 78(1) January 1990.
[13] Swanson, Don R. & Smalheiser, Neil R. "Implicit Text Linkages Between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery." Library Trends, Vol. 48, No. 1, Summer 1999, pp. 48-59.
[14] Swanson, Don R. "Online Search for Logically-Related Noninteractive Medical Literatures: a Systematic Trial-and-Error Strategy." Journal of the American Society for Information Science. 40 (5):356-358. 1989.
[15] Swanson, Don R. & Smalheiser, Neil R. "An Interactive System for Finding Complementary Literatures: a Stimulus to Scientific Discovery". Artificial Intelligence 91 (1997), pp. 183-203.
[16] Swanson, Don R., et al. ARROWSMITH web site
http://d-swanson.uchicago.edu/
[17] Gordon, Michael D. & Lindsay, Robert K. "Toward Discovery Support Systems: a Replication, Re-

Examination, and Extension of Swanson's Work on Literature-Based Discovery of a Connection between Raynaud's and Fish Oil". Journal of the American Society for Information Science. 47(2): 116-128. 1996.
[18] Garvey, William D., Communication: the Essence of Science. Oxford: Pergmon Press, 1979.
[19] Kuhn, Thomas S. The Structure of Scientific Revolutions, Second Edition. Volume 2, Number 2, "The Structure of Scientific Revolutions" in the International Encyclopedia of Unified Science, Otto Neurath, Editor-in-Chief. University of Chicago Press, Chicago IL, 1970. (First Edition, 1962.)
[20] Popper, Karl R. Conjectures and Refutations: the Growth of Scientific Knowledge. Basic Books, Inc., New York, 1963.
[21] Palmer, Carole L. Structures and Strategies of Interdisciplinary Science. JASIS 50(3):242-253, 1999.
[22] Palmer, Carole L. *Practices and Conditions of Boundary Crossing Research Work: A Study of Scientists at an Interdisciplinary Institute*. Ph.D. Thesis, UIUC GS-LIS. February 1996.
[23] Braam, Robert R. *Mapping of science : foci of intellectual interest in scientific literature* / Robert R. Braam. Leiden, Netherlands : DSWO Press, University of Leiden, 1991.
[24] Haythornthwaite, Caroline. Social Network Analysis: An Approach and Technique for the Study of Information Exchange. Library and Information Science Research, 18 (4), 323-342, 1996.
[25] Wasserman, Stanley & Faust, Katherine. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.
[26] Braam, Robert R. Mapping of science : Critical Elaboration and New Approaches / R.R. Braam, H. F. Moed, A.F.J. van Raan. Leiden, Netherlands : DSWO Press, University of Leiden, 1991.
[27] Sandstrom, Pamela Effrein, "An Optimal Foraging Approach to Information Seeking and Use", Library Quarterly, vol. 64, no. 4, pp. 414-449, 1994.
[28] Sandstrom, Pamela Effrein, "Scholars as Subsistence Foragers", Bulletin of the American Society for Information Science, Vol. 25 #3, February/March 1999.
http://www.asis.org/Bulletin/Feb-99/sandstrom.html
[29] Von Neumann, John & Morgenstern, Oskar. Theory of Games and Economic Behavior. Princeton University Press, London. Second Edition, 1947.
[30] Spasser, Mark A. "The Enacted Fate of Undiscovered Public Knowledge". Journal of the American Society for Information Science. 48(8):707-717, 1997.
[31] Schatz, Bruce; Mischo, William; Cole, Timothy; Bishop, Ann; Harum, Susan; Johnson, Eric; Neumann, Laura; Chen, Hsinchun; Ng, Dorbin. "Federated Search of Scientific Literature", IEEE Computer, Special Issue on Digital Libraries, 32: 51-59, February 1999.
[32] National Center for Biotechnology Information, PubMed Overview, as of August 27, 2001.
http://www.ncbi.nlm.nih.gov/entrez/journals/loftext_noprov.html
[33] Lancaster, F. Wilf, Palmer, Carole L., Jackson, Larry S., Spasser, Mark A., Dubin, David. LBD/ UPK Supercomputing Project homepage.
http://alexia.lis.uiuc.edu/gslis/research/auto-methods.html